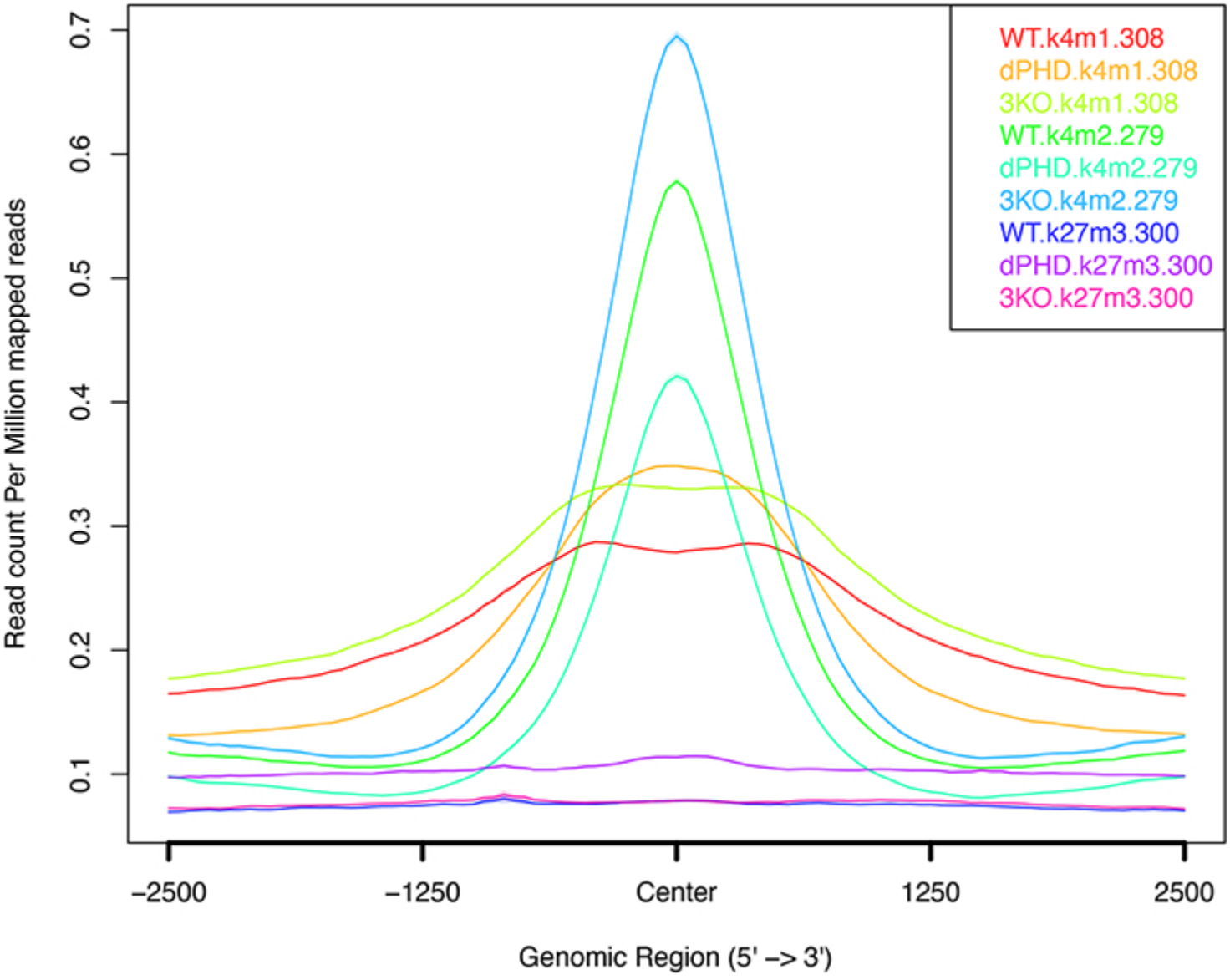
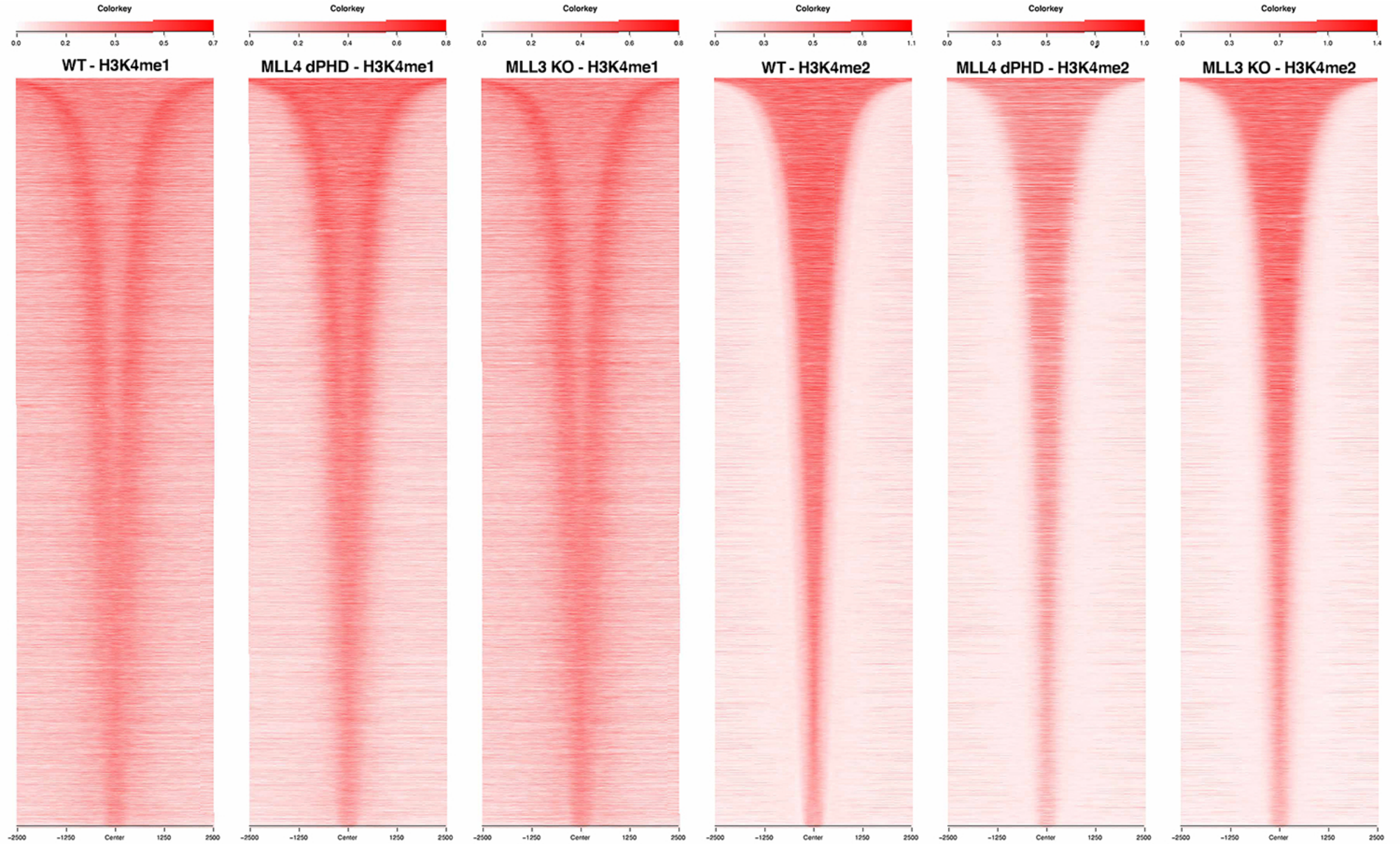
NGSPLOT PIPELINE

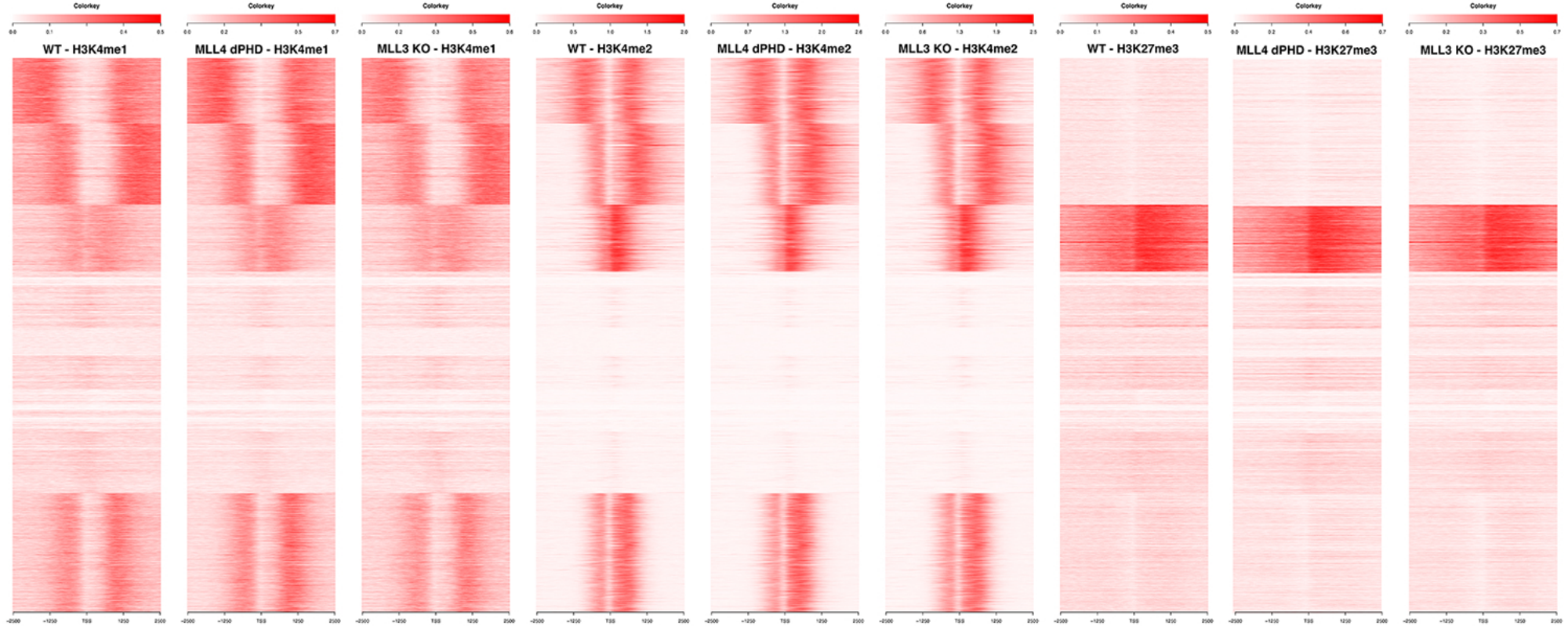
CLAYTON K. COLLINGS, PHD

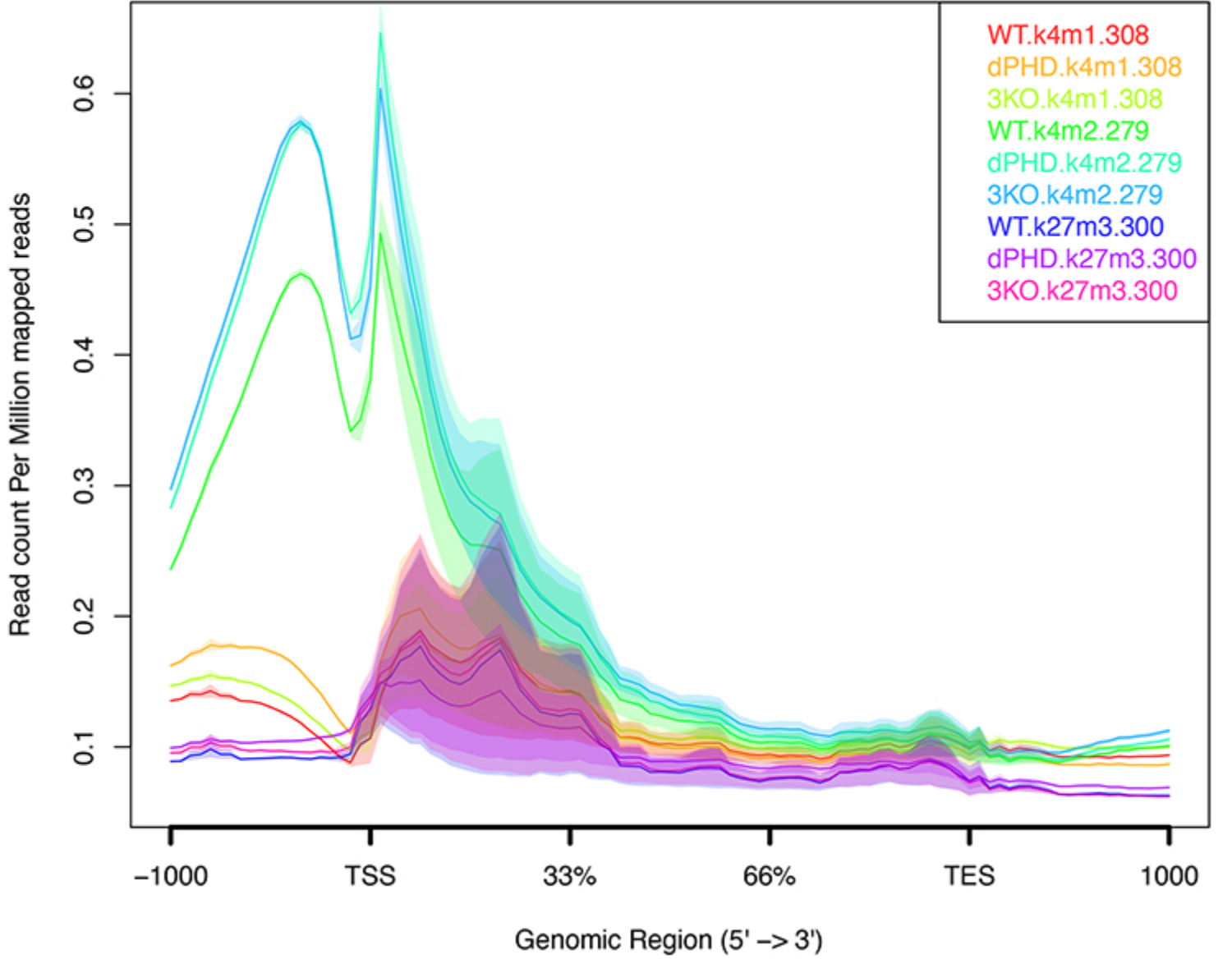
PIPELINE FUNCTION

- Generates metaplots, metagene plots, and heatmaps of occupancy levels from ChIP-seq data
- Compares treated versus control data at peaks, TSSs, and along gene bodies by plotting log2 fold changes in heatmaps
- Partitions these log2-fold-change heatmaps into groups by k-means clustering
- Annotates peaks and performs motif analysis on data clusters
- Plots RNA-seq log2 fold changes along side ChIP-seq data
- Performs GO analysis on genes in each cluster


WT.k4m1.308
dPHD.k4m1.308
3KO.k4m1.308
WT.k4m2.279
dPHD.k4m2.279
3KO.k4m2.279
WT.k27m3.300
dPHD.k27m3.300
3KO.k27m3.300
Read count Per Million mapped reads
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−2500
−1250
Center
1250
2500
Genomic Region (5' −> 3')

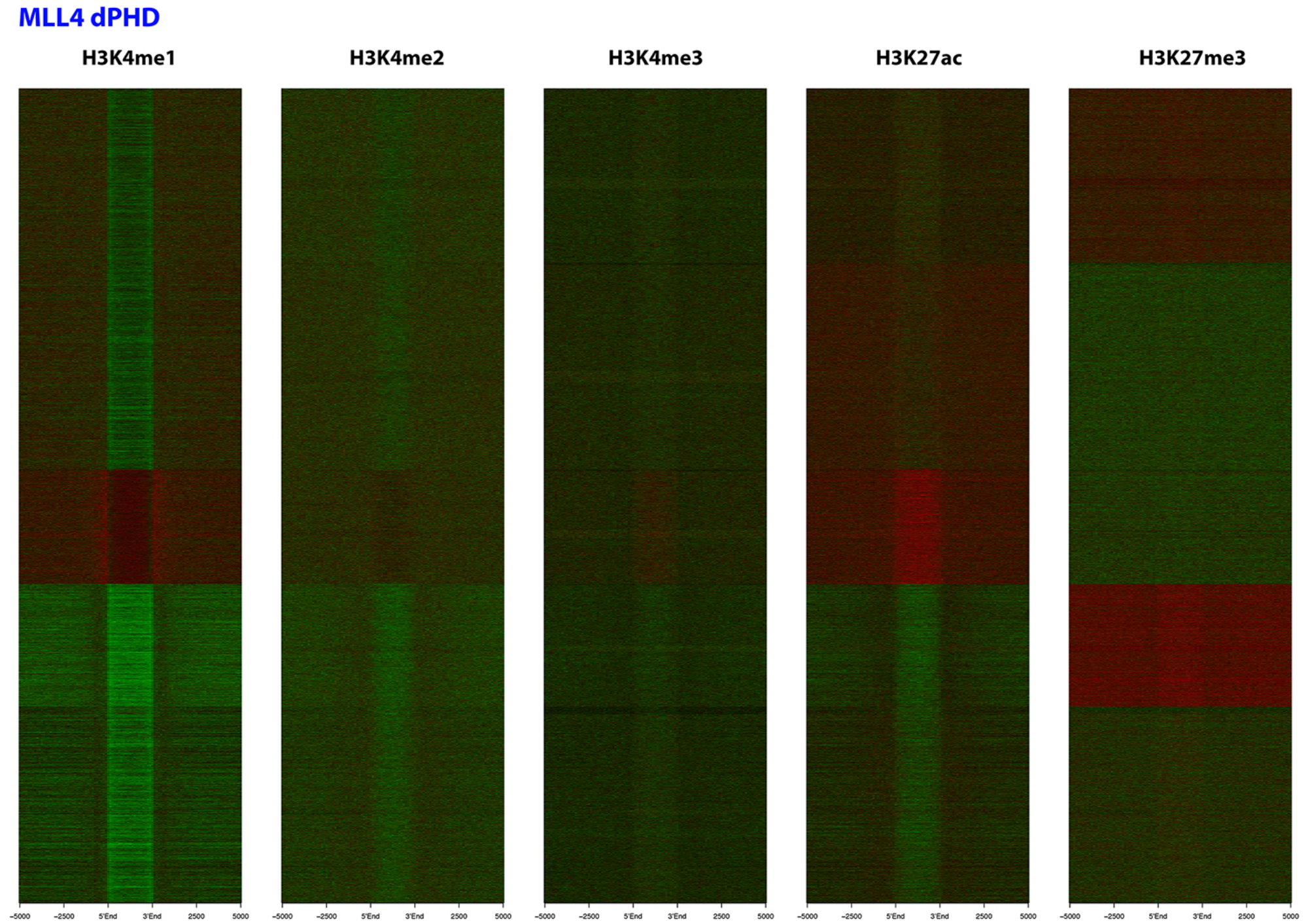

Colorkey
0.0 0.2 0.3 0.5 0.7
WT - H3K4me1
Colorkey
0.0 0.2 0.4 0.6 0.8
MLL4 dPHD - H3K4me1
Colorkey
0.0 0.2 0.4 0.6 0.8
MLL3 KO - H3K4me1
Colorkey
0.0 0.3 0.5 0.8 1.1
WT - H3K4me2
Colorkey
0.0 0.3 0.5 0.8 1.0
MLL4 dPHD - H3K4me2
Colorkey
0.0 0.3 0.7 1.0 1.4
MLL3 KO - H3K4me2
-2500 -1250 Center 1250 2500


Colorkey
WT - H3K4me1
MLL4 dPHD - H3K4me1
MLL3 KO - H3K4me1
WT - H3K4me2
MLL4 dPHD - H3K4me2
MLL3 KO - H3K4me2
WT - H3K27me3
MLL4 dPHD - H3K27me3
MLL3 KO - H3K27me3
−2500
−1250
TSS
1250
2500

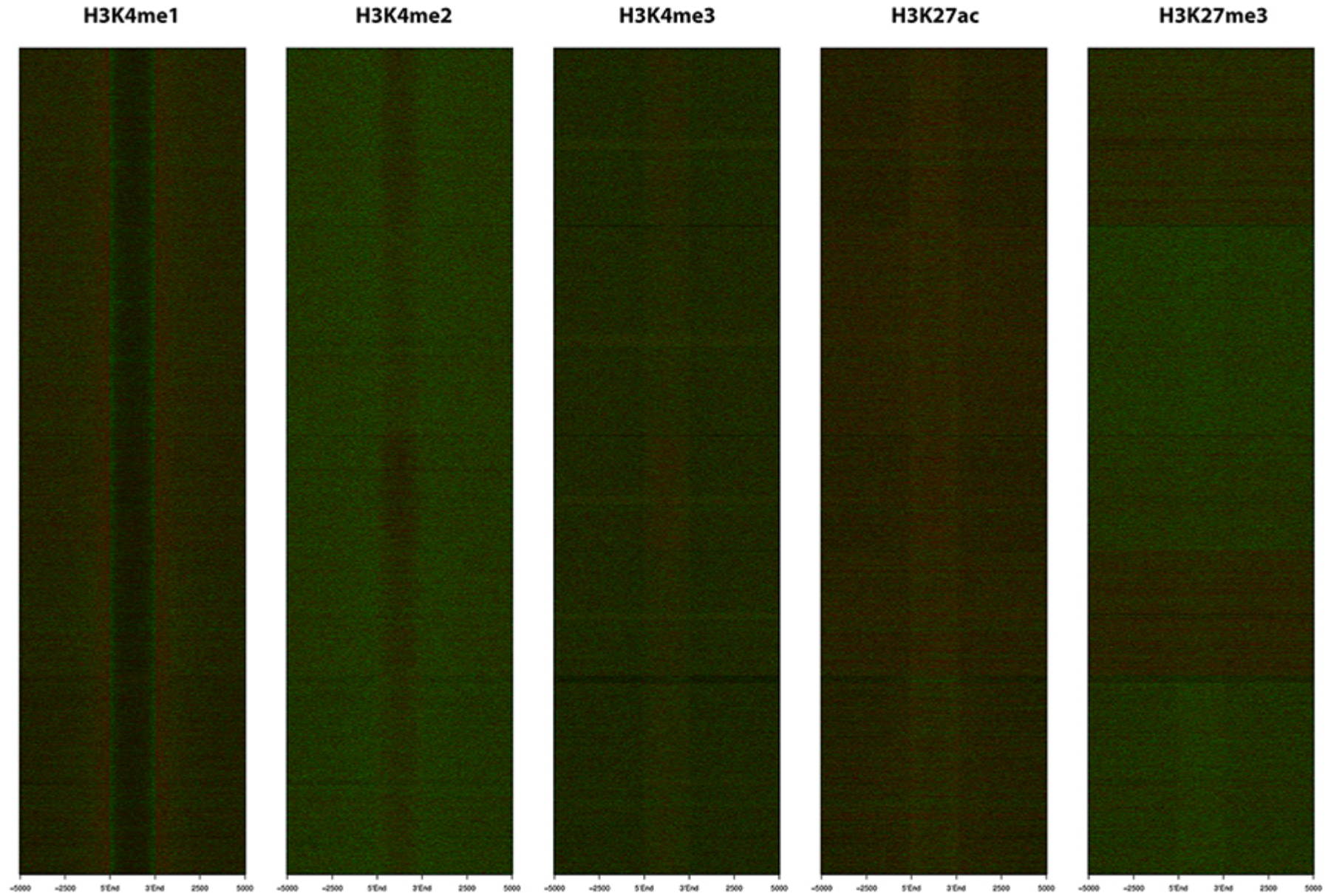

WT.k4m1.308
dPHD.k4m1.308
3KO.k4m1.308
WT.k4m2.279
dPHD.k4m2.279
3KO.k4m2.279
WT.k27m3.300
dPHD.k27m3.300
3KO.k27m3.300
Read count Per Million mapped reads
0.1
0.2
0.3
0.4
0.5
0.6
−1000
TSS
33%
66%
TES
1000
Genomic Region (5' −> 3')

PIPELINE FUNCTION

- Generates metaplots, metagene plots, and heatmaps of occupancy levels from ChIP-seq data
- Compares treated versus control data at peaks, TSSs, and along gene bodies by plotting log2 fold changes in heatmaps
- Partitions these log2-fold-change heatmaps into groups by k-means clustering
- Annotates peaks and performs motif analysis on data clusters
- Plots RNA-seq log2 fold changes along side ChIP-seq data
- Performs GO analysis on genes in each cluster


MLL4 dPHD
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K27me3
−5000
−2500
5'End
3'End
2500
5000

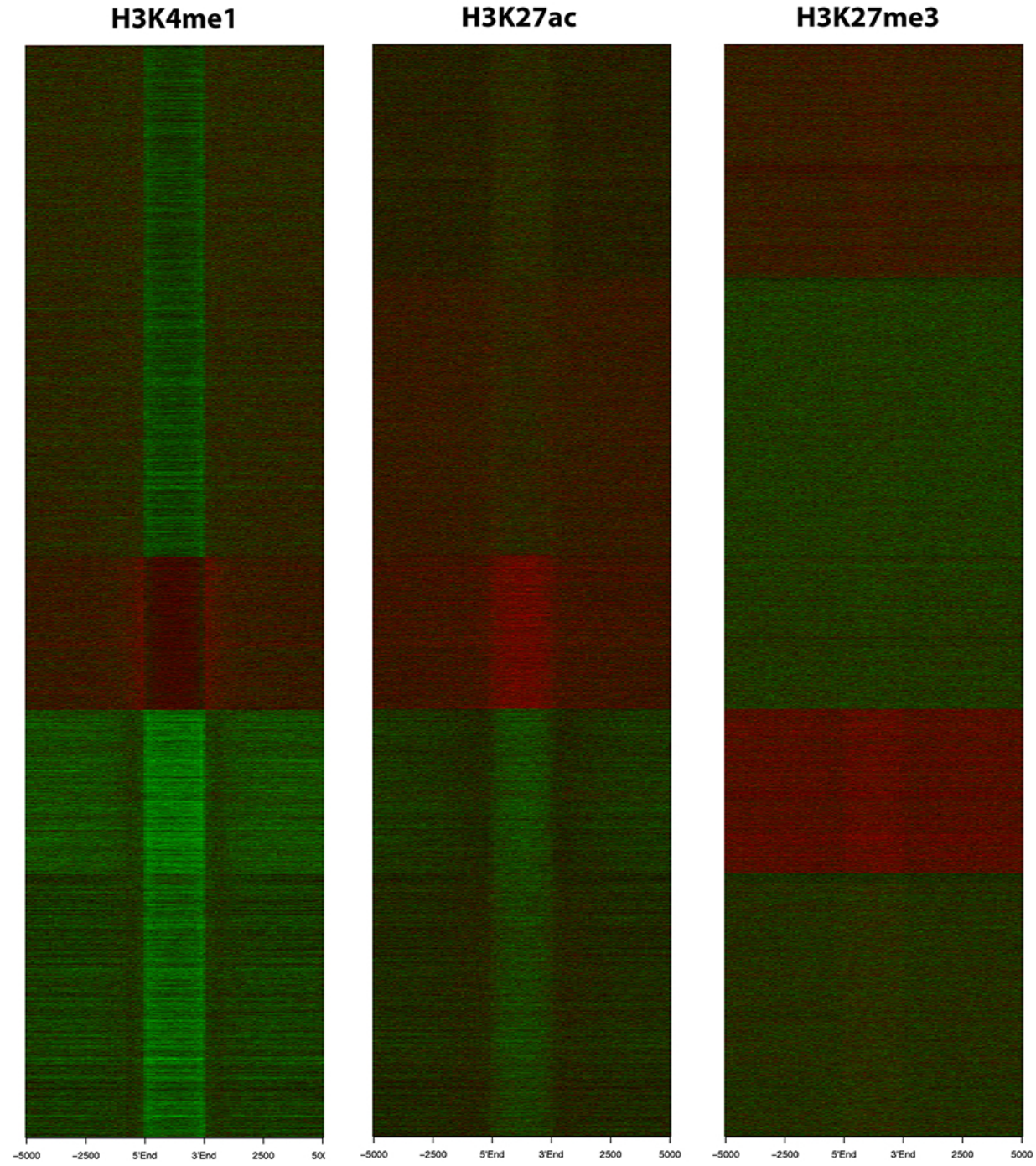
MLL3 KO


H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K27me3
-5000
-2500
5'End
3'End
2500
5000

PIPELINE FUNCTION

- Generates metaplots, metagene plots, and heatmaps of occupancy levels from ChIP-seq data
- Compares treated versus control data at peaks, TSSs, and along gene bodies by plotting log2 fold changes in heatmaps
- Partitions these log2-fold-change heatmaps into groups by k-means clustering
- Annotates peaks and performs motif analysis on data clusters
- Plots RNA-seq log2 fold changes along side ChIP-seq data
- Performs GO analysis on genes in each cluster

MLL4 dPHD

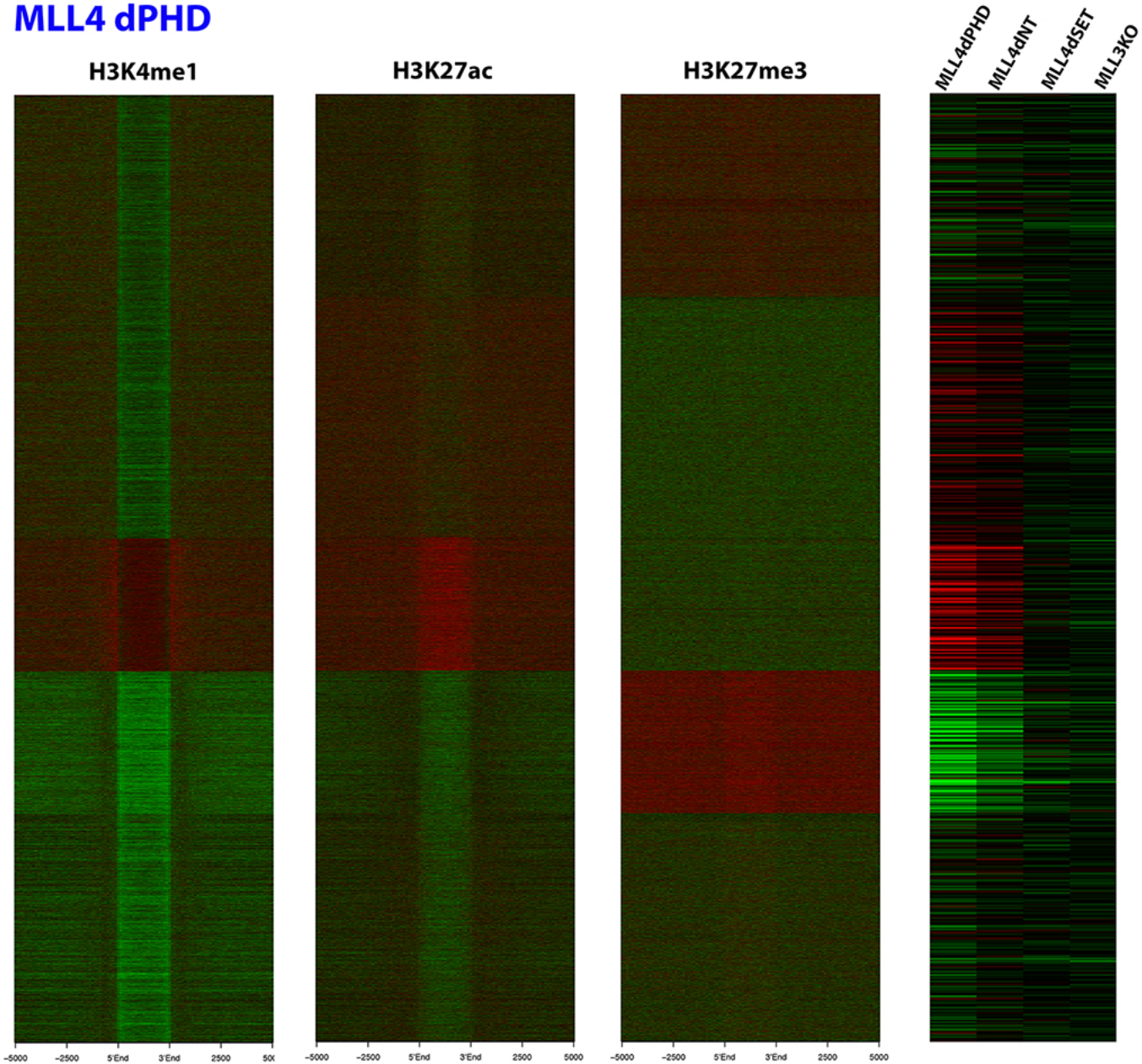


For all peaks or for individual clusters, such as clusters 3 and 4, the pipeline automatically sets up codes to perform various analyses.

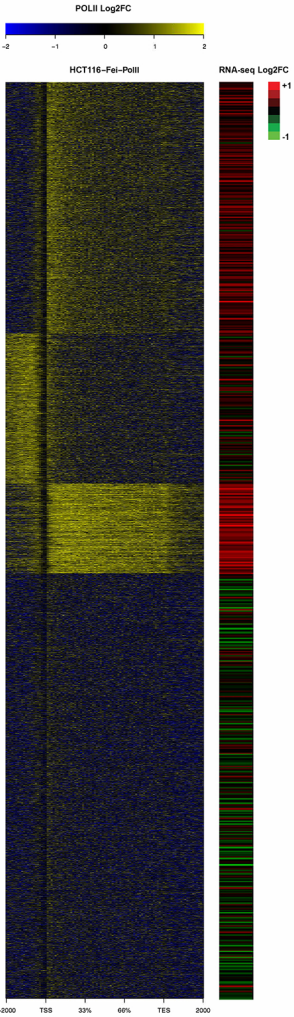
1. **Peak Annotation**
2. **Motif Analysis**
3. **Nearest Gene Ontology**
4. **Genome Ontology Analysis**

PIPELINE FUNCTION

- Generates metaplots, metagene plots, and heatmaps of occupancy levels from ChIP-seq data
- Compares treated versus control data at peaks, TSSs, and along gene bodies by plotting log2 fold changes in heatmaps
- Partitions these log2-fold-change heatmaps into groups by k-means clustering
- Annotates peaks and performs motif analysis on data clusters
- Plots RNA-seq log2 fold changes along side ChIP-seq data
- Performs GO analysis on genes in each cluster



Additionally, the pipeline can plot the log fold changes of the nearest genes to peaks and perform GO analysis for genes in clusters as well as gene subsets within these clusters using p-value cutoffs from edgeR.



The pipeline can also peform metagene analyses, linking changes in gene expression from RNA-seq data with changes in differential occupancy from ChIPseq data.

For example, differential occupancy of Pol2 after PAF1 knockdown is displayed in the blue and yellow metagene heatmap for all genes.

Corresponding log fold changes in gene expression are plotted along the side in green and red. These log fold changes are typically derived from the edgeR output generated from our standard data processing pipeline.

GO analysis can be performed on gene clusters of interest, and p-value cutoffs can be implemented to obtain gene subsets within these clusters.

PIPELINE PURPOSE AND USAGE

This pipeline will save time, prevent mistakes, and preserve this data analysis approach for future bioinformaticians in this lab.

With one command, one can generate all of the analyses mentioned in this presentation.

The following slides contain summarized and detailed descriptions of inputs, outputs, options, and instructions for the pipeline.

PIPELINE INPUTS / OPTIONS SUMMARY

- sample list (mapped ChIP-seq files)
- comparison list (two-sample comparisons)
- peak list (files with peak starts and stops)
- order of peaks/genes in heatmap (high-to-low sort or k-means)
- heatmap scales, color, cluster number
- gene list (all genes or user-provided gene list)
- edgeR list (contains RNA-seq log2FCs and p-values)

PRIMARY INPUTS/OPTIONS			
option	*******	**description**	**default**
outputDirectory	o	output directory location	./
outputShell	os	output shell script name/location	./makeNGSplots.sh
bamDirectory	bam	will use bam files in directory instead of sample list	
sampleList	s	<sampleName>\t<pathToSample>	
bedList	b	<bedName>\t<pathToBed>	
comparisons	c	<compName>\t<pathToTreated:pathToControl>	
comparisonBedList	cb	<bedName>\t<pathToBed>	
ngsPlotFilesScriptsDirectory	ngs	pipeline directory	/projects/b1025/NGSplotPipeline/NGSplotFilesScripts
homerShellScript	hss	homer shell script name/location	./homerShellScript.sh
homerShellScript2	hss2	homer shell script name/location	./homerShellScript2.sh
assembly	g	genome reference symbol	
numProcessors	p	number of processors	4
fragmentLength	fl	fragment length	150

notes:
bamDirectory, sampleList, or comparisons must be specified or pipeline will not run
.bai files must be present

OCCUPANCY PLOT OPTIONS			
option	***	**description**	**default**
bedLength	bl	window around feature center or start/stop	2500
sortBed	sb	sorts heatmap by feature width	1
centerBed	cenb	aligns data to feature centers	1
useEqualPeakWidthforBedList	epw	makes all features the same width in heatmap	0
bedOrder	bo	order of features in heatmap	none
bedClusters	bc	when bedOrder=km, bc = # of clusters	5
heatmapScale	hs	rpm scale, use <min,max> i.e. 0,1	auto
heatmapColor	hc	color of heatmaps	red
verticalLines	vl	vertical line presence in average plots	0
lineWidth	lw	line thickness in average plots	1
runTSS	rt	aligns data to transcription start sites	0
runGenebody	rgb	makes metagene plots	0
runExon	re	aligns data to exons	0
lengthTSS	tl	distance from TSS for aligning data	2500
lengthGenebody	gbl	distance from TSS and TTS for aligning data	1000
lengthExon	el	distance from exon for aligning data	500
geneList	gl	user provided gene list	
geneOrder	go	order of genes in heatmaps	km
geneClusters	gc	when geneOrder=km, gc = # of clusters	5

notes:
if clustered occupancy heatmaps around bed positions are desired, set bedOrder=km
if clustered occupancy heatmaps around bed positions are desired, set useEqualPeakWidthforBedList=1, sortBed=0, & centerBed=0
if clustered occupancy heatmaps around bed positions are desired, recommend bedLength=1000
if gene order of user-provided gene list is desired to be maintained, set geneOrder=none
if bedOrder is desired to be maintained, set useEqualPeakWidthforBedList=1 OR centerBed=1, sortBed=0, & bedOrder=none

COMPARISON PLOT OPTIONS			
option	***	**description**	**default**
comparisonBedLength	cbl	window around feature center or start/stop	bedLength
comparisonSortBed	csb	sorts comparison heatmap by feature width	0
comparisonCenterBed	ccenb	aligns data to feature centers	0
useEqualPeakWidthforComparisonBedList	cepw	makes all features the same width in heatmap	1
comparisonBedOrder	cbo	order of features in heatmap	km
comparisonBedClusters	cbc	when comparisonBedOrder=km, cbc = # of clusters	5
runComparisonTSS	rct	aligns log2 FC data to TSSs	0
runComparisonGenebody	rcgb	makes metagene plots with log2 FC values	0
runComparisonExon	rce	aligns log2 FC data to exons	0
comparisonGeneList	cgl	user provided gene list for comparisons	
comparisonGeneOrder	cgo	order of genes in comparison heatmap	km
comparisonGeneClusters	cgc	number of clusters of genes in heatmap	5
comparisonHeatmapScale	chs	log2 FC scale, use <min,max> i.e. -2,2	auto
comparisonHeatmapColors	chc	tri color specification for comparison heatmap	skyblue:black:yellow
comparisonCD	ccd	heatmap color distribution	0.6
edgeRlist	erl	<edgeRname>\t<pathToEdgeRfile>	
edgeRpv	erp	p value cutoff for edgeR	0.05

notes:
recommend comparisonBedOrder=km and comparisonGeneOrder=km
if comparison gene order of user-provided gene list is desired to be maintained, set comparisonGeneOrder=none
if comparison bed order is desired to be maintained, set useEqualPeakWidthforComparisonBedList=1 OR comparisonCenterBed=1, comparisonSortBed=0, & comparisonBedOrder=none
edgeR files provide logFC values in cdt files that line up with features or genes in comparison heatmaps
edgeR p value is used to get lists of significantly upregulated or downregulated genes that overlap genes (or nearest genes) in clusters